Prediction of Lymph Node Metastasis in Colorectal Cancer using Intraoperative Fluorescence Multi-modal Imaging

Xiaobo Zhu, He Sun, Yuhan Wang, Gang Hu, Lizhi Shao, Song Zhang, Fucheng Liu, Chongwei Chi, Kunshan He, Jianqiang Tang, Yu An,*Member, IEEE*, Jie Tian, *Fellow, IEEE*, and Zhenyu Liu, *Member, IEEE*

***Abstract*— The diagnosis of lymph node metastasis (LNM) is essential for colorectal cancer (CRC) treatment. The primary method of identifying LNM is to perform frozen sections and pathologic analysis, but this method is labor-intensive and time-consuming. Therefore, combining intraoperative fluorescence imaging with deep learning (DL) methods can improve efficiency. The majority of recent studies only analyze uni-modal fluorescence imaging, which provides less semantic information. In this work, we mainly established a multi-modal fluorescence imaging feature fusion prediction (MFI-FFP) model combining white light, fluorescence, and pseudo-color imaging of lymph nodes for LNM prediction. Firstly, based on the properties of various modal imaging, distinct feature extraction networks are chosen for feature extraction, which could significantly enhance the complementarity of various modal information. Secondly, the multi-modal feature fusion (MFF) module, which combines global and local information, is designed to fuse the extracted features. Furthermore, a novel loss function is formulated to tackle the issue of imbalanced samples, challenges in differentiating samples, and enhancing sample variety. Lastly, the experiments show that the model has a higher area under the receiver operating characteristic (ROC) curve (AUC), accuracy (ACC), and F1 score than the uni-modal and bi-modal models and has a better performance compared to other efficient image classification networks. Our study demonstrates that the MFI-FFP model has the potential to help doctors predict LNM and shows its promise in medical image analysis.**

***Index Terms*— Deep learning, multi-modal imaging, intraoperative fluorescence imaging, colorectal cancer, lymph node metastasis prediction.**

Manuscript received 22 May 2024. Our code and model can be available at: *https://github.com/Emibobo/MFI-FFP*. This work is supported by the National Natural Science Foundation of China under Grant Nos. 62333022, 92259301, 81930053, and 62027901, the National Key R&D Program of China under Grant Nos. 2021YFF1201003, 2023YFC3402800, and 2022YFC2407400, and the Beijing Natural Science Foundation under Grant Nos. JQ23034, L234056, L232097, L222054, and 4232058. (Xiaobo Zhu, He Sun, and Yuhan Wang contribute equally to the article. Corresponding authors: Jianqiang Tang, Yu An, Jie Tian, and Zhenyu Liu.)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the study, which was approved by the Ethics Committee of Beijing Fengtai Hospital and Cancer Hospital, Chinese Academy of Medical Sciences.

Xiaobo Zhu, Lizhi Shao, Song Zhang, Chongwei Chi, and Kunshan He are with the CAS Key Laboratory of Molecular Imaging and Beijing Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zhuxiaobo2023@ia.ac.cn; lizhi.shao@ia.ac.cn; zhangsong2021@ia.ac.cn; chongwei.chi@digipmc.com; kunshan.he@digipmc.com).

He Sun and Yu An are with Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, School of Engineering Medicine, Beihang University, Beijing 100191, China (e-mail: hesun@buaa.edu.cn; yuan1989@buaa.edu.cn).

Fucheng Liu is with the Department of General Surgery, Beijing Fengtai Hospital, Beijing 100071, China (e-mail: sqzr_ren@sina.com).

Yuhan Wang, Gang Hu, and Jianqiang Tang are with the Department of Colorectal Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China (e-mail: wyh06112023@163.com; doctorhuhu@126.com; doc_tjq@hotmail.com).

Jie Tian and Zhenyu Liu are with the CAS Key Laboratory of Molecular Imaging and Beijing Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with National Key Laboratory of Kidney Diseases, Beijing, 100853, China (e-mail: jie.tian@ia.ac.cn; zhenyu.liu@ia.ac.cn).

I. Introduction

COLORECTAL cancer (CRC) is the third most commonly diagnosed cancer and the third most common cause of cancer-related deaths worldwide [1]. The incidence of CRC has increased in recent years, and it accounts for about 10% of all cancers [2]. Complete resection of the primary tumor and regional lymph nodes is considered the most critical aspect of CRC treatment [3]. The lymph node metastasis (LNM) status significantly influences treatment decisions, including the choice of preoperative neoadjuvant radiotherapy versus surgery [4]. Preoperative conventional imaging methods such as computed tomography (CT) and magnetic resonance imaging (MRI) have some limitations in detecting LNM, such as the risk of radiation, low diagnostic efficiency that requires expert interpretation, etc., and need to be combined with other techniques to confirm the diagnosis [5]–[7]. Although, the utilization of lymph node biopsy can more accurately assess LNM; however, it leads to an expensive investment of time, labor, and other costs [8].

In CRC, Zhao et al. explored potential genomic phenotypes associated with deep learning (DL) features that facilitate

LNM prediction [9]. Lennard et al. enhanced the prediction performance by utilizing a DL technique to analyze histological whole slide images (WSIs) to predict LNM in CRC [10]. Furthermore, Li and colleagues proposed improving the accuracy of LNM identification through the application of MRI in conjunction with a pre-trained Inception-v3 model on transfer learning [11]. However, the majority of studies on the prediction of LNM in CRC use either preoperative CT, MRI, or pathology imaging, or combine tissue genetic information with deep learning methods for prediction, which significantly raises implementation costs and reduces overall efficiency. Furthermore, these preoperative imaging techniques are incapable of visualizing the complete architecture of the lymph node, and their intraoperative utility is significantly restricted, leading to diminished detection efficiency and accuracy [12]. Nevertheless, intraoperative fluorescence navigation systems enable both surgical navigation during surgery and real-time imaging of multi-modal fluorescence images to visualize tissues [13]. Due to the quick capture time and affordable nature of intraoperative multi-modal fluorescence imaging, the analysis of multi-modal fluorescence imaging not only decreases the time and cost of labor but also has the potential to assist in making clinical decisions during surgery. Hence, we propose integrating DL with intraoperative fluorescent multi-modal imaging to enhance the efficiency and accuracy of intraoperative LNM prediction significantly.

Indocyanine green (ICG) is considered to be more effective in CRC lymph node identification since ICG penetrates relatively deeply into the tissue of the colon compared to other fluorescent tracers [14]. When injected intravascularly, ICG binds to proteins and is transported through the lymphatic system, normally draining to the closest lymph node in 15 minutes or less. It is becoming more typical to employ near-infrared (NIR) fluorescence imaging (FI) to guide surgical guidance following ICG injection [15], [16]. Fluorescence-guided surgery offers advantages such as real-time imaging, high visualization, enhanced contrast, non-invasiveness, precise navigation, and broad applicability, all of which contribute to improved surgical safety, accuracy, and success rates. Typical fluorescence surgical imaging is available in three modes, white light imaging (WLI), fluorescence imaging (FI), and pseudo-color imaging (PCI). Most of the existing studies use DL to analyze FI, mainly for cancer diagnosis or image enhancement, etc., while few studies have been conducted to predict LNM.

Compared to machine learning-based radiomics approaches, DL offers the advantage of automatically learning hierarchical features from raw data, eliminating the need for manual feature extraction, and often achieving superior performance in tasks such as image classification, segmentation, and pattern recognition [17]. In the field of medical image analysis, DL has been widely utilized for analyzing multi-modal images, including CT and MRI. However, most multi-modal medical imaging analysis methods employ the same feature extraction network, which is unable to take into account the specificity of distinct imaging features. In addition, most of the methods for multi-modal medical imaging feature fusion in the intermediate stage of the model directly integrate the features by dimensionally concatenating them or utilizing the Transformer structure [18], but this is not able to focus on the feature information at both the global and local levels.

Most studies have increasingly incorporated DL methods to process intraoperative FI for targeted medical tasks. Shen et al. developed an FL-CNN model to extract relevant information from glioma FI, enhancing the precision of intraoperative glioma diagnosis to assist surgeons in accurately identifying and removing gliomas [19]. Cahill et al. proposed a method combining artificial intelligence (AI) with intraoperative FI video analysis, which was utilized to improve the accuracy of tumor identification for intraoperative tissue classification of colorectal cancer [20]. The growing number of researchers leveraging AI to analyze intraoperative fluorescence imaging or video highlights the advantages of fluorescence-based surgical guidance [21]–[23]. Additionally, Xiao et al. developed a method for intraoperative glioma grading by combining various imaging features and introduced DLS-DARTS to fuse multi-modal near-infrared (NIR) fluorescence imaging, resulting in improved predictive outcomes [24]. Multi-modal medical imaging learning can take advantage of data from different medical imaging modalities to achieve comprehensive analysis and diagnosis. Most of the studies analyzed FI, while a few combined FI and WLI but lacked the analysis of PCI. Hence, three intraoperative modality imaging types, WLI, FI, and PCI, are utilized in our study to increase the information on lymph node features in CRC, which improves the prediction performance.

Here, we propose a multi-modal imaging learning framework, called the multi-modal fluorescence imaging feature fusion prediction (MFI-FFP) model, to extract and integrate multi-modal intraoperative fluorescence information, including the three modalities of WLI, FI, and PCI, for LNM prediction in CRC. In this study, a CRC lymph node dataset is collected and prepared specifically for the three modalities of WLI, the first NIR fluorescence window (NIR-I, 700-900 nm) FI, and PCI. First, we design a feature extraction framework comprising three pre-trained branches to extract lymph node features from the tri-modal images: the Inverse Residual Network branch for white light image features, the Residual Network branch for fluorescence image features, and the Transformer Network branch for pseudo-color image features. Subsequently, the integration of global and local feature information from the three modal data is done by a multi-modal feature fusion (MFF) module, which then produces the fusion lymph node features. Moreover, LNM in cancer patients typically exhibit a more pronounced benign-malignant imbalance, with the benign-malignant ratio escalating to 10:1 in extreme instances [25]–[27]. Therefore, a novel loss function is tailored to the data characteristics to address the issue of class imbalance and the challenge of distinguishing between classes in the dataset. The main contributions of our work are summarized as follows:

- To fully utilize the feature information from each imaging modality, we select three specialized feature extraction branches for the tri-modal images. This approach enhances the model's ability to capture and characterize features specific to each modality, and experimental results

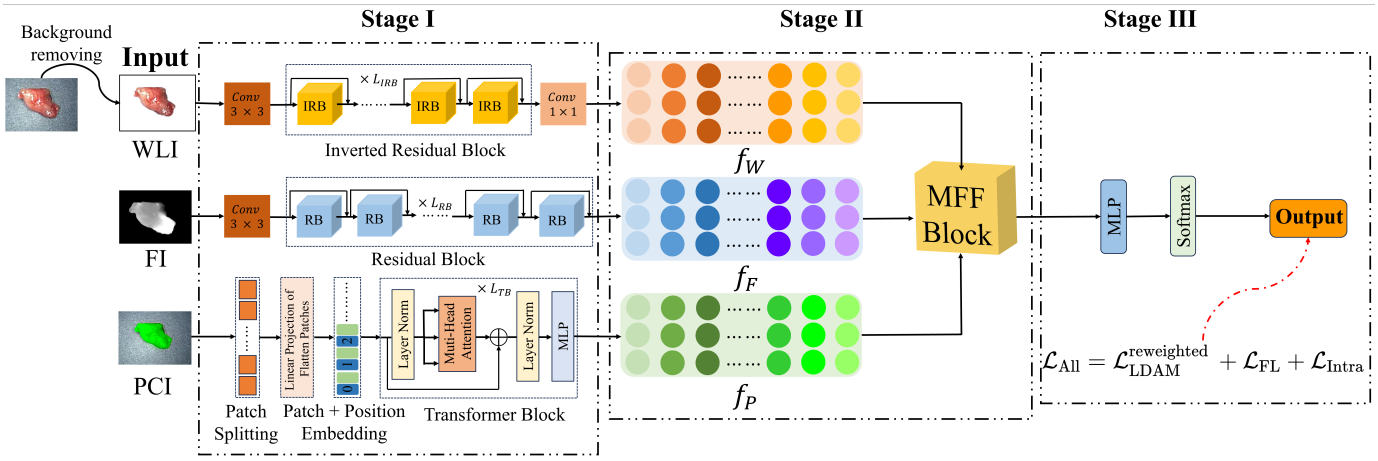

Fig. 1. The overall framework of the MFI-FFP Model. Stage I. The features of the three modal data are extracted using the inverted residual, residual, and Transformer structures of the feature extraction branches, respectively. Stage II. The three extracted modal features are fused at the global and local levels through the MFF module. Stage III. Lymph node metastasis prediction based on fused features.

confirm the effectiveness of this design.

- To comprehensively utilize the complementary information contained in different modal imaging and provide more comprehensive and accurate imaging characterization, we design a multi-modal feature fusion (MFF) module to enhance the model's ability to combine multi-modal imaging information for decision-making.
- We propose a novel intra-class loss function based on the combination of label-distribution-aware margin (LDAM) loss and Focal loss, which increases the diversity of the data while reducing the impact of the sample imbalance problem.

II. METHODOLOGY

A. Multi-modal Imaging Feature Extraction Branches

As demonstrated in Fig. 1, three feature extraction branches are built at the beginning of the entire MFI-FFP model to extract WLI, FI, and PCI image features. The advantage of extracting multiple modal images utilizing distinct feature extraction branches is that the feature information of each imaging modality can be fully utilized, potentially enhancing the model's characterization capacity. To ensure that the primary feature information is retrieved with the least amount of loss, we utilize a background removal method for the WLI data to eliminate superfluous background interference noise. Given that white light imaging (WLI) is characterized by high pixel density and numerous features, we employ an inverted residual structure to efficiently capture spatial details and structural features. This module enhances computational efficiency by reducing the number of operations in the bottleneck layer while preserving critical feature representations, enabling lightweight feature extraction. This makes it well-suited for WLI, which typically demands high-resolution and detailed spatial information. The inverted structure is particularly advantageous for processing high-dimensional images like WLI, as it balances accuracy with computational complexity.

In contrast, a conventional ResNet is used for fluorescence imaging, which has three-channel values resembling the characteristics of grayscale images. ResNet is effective at capturing both fine and coarse features across different depths. Since the spatial variability in fluorescence imaging is generally lower than that in WLI, a standard ResNet is sufficient for extracting the necessary hierarchical features without the need for additional computational optimizations.

PCI data, which includes tissue size, edge feature, and fluorescence intensity information is the result of mapping fluorescence feature information on tissue imaging. To avoid the WLI data losing some edge information during background removal, no background removal procedure is applied to the PCI data in this study. The Vision Transformer (ViT) structure is used for pseudo-color imaging (PCI) due to its ability to capture long-range dependencies within the data. Unlike convolutional networks, the ViT model establishes connections between local and global features, allowing it to focus on multiple parts of the image simultaneously. This capability enables effective capture of both local and global relationships in PCI, which is essential for accurate feature extraction. The extracted features from PCI can effectively complement the WLI and FL features, enhancing the overall feature representation and model performance.

In the feature extraction network branches, the WLI, FI, and PCI sample spaces are defined as $X_W = \{x_W^1, x_W^2, \ldots, x_W^n\}$, $X_F = \{x_F^1, x_F^2, \ldots, x_F^n\}$, and $X_P = \{x_P^1, x_P^2, \ldots, x_P^n\}$, respectively, and x_W^i, x_F^i, and $x_P^i \in \mathbb{R}^{C\times H\times W}$. The corresponding inverse residual, residual, and ViT feature extraction branches are defined as functions $\mathbf{F_{IR}}$, $\mathbf{F_R}$, and $\mathbf{F_{ViT}}$, respectively. Thus, the procedure for extracting features is as follows in (1) to (3).

$$f_W^i = \mathbf{F_{IR}}(x_W^i) \tag{1}$$

$$f_F^i = \mathbf{F_R}(x_F^i) \tag{2}$$

$$f_P^i = \mathbf{F_{ViT}}(x_P^i) \tag{3}$$

where f_W^i, f_F^i, and f_P^i are the three modal imaging features corresponding to the lymph node with index i, respectively. In

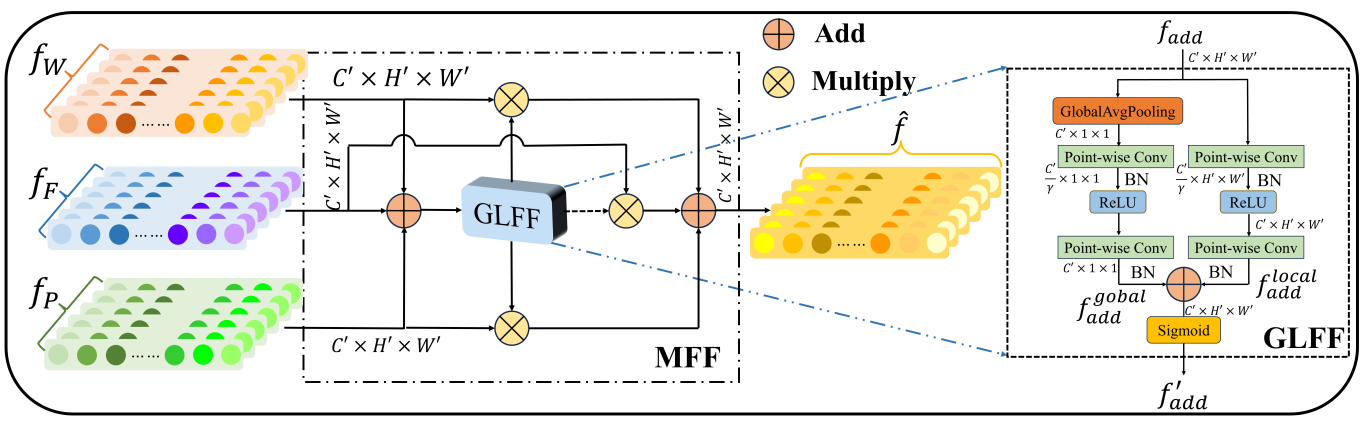

Fig. 2. The overall structure of the MFF module. The extracted features from the three modal data are fused with the global and local hierarchical features through the GLFF module and a series of weighting operations are performed with the original features to produce the final fused features.

this case, the ViT model is calculated as follows in (4) to (7).

$$z_0 = \left[x_{\text{class}}\,; x_p^1 E; x_p^2 E \cdots x_p^N E\right], E \in \mathbb{R}^{(P^2 \cdot C) \times D} \tag{4}$$

$$z'_\ell = \text{MSA}\left(\text{LN}\left(\text{z}_{\ell-1}\right)\right) + \text{z}_{\ell-1}, \ell \in [1, 2, \cdots, L] \tag{5}$$

$$\text{z}_\ell = \text{MLP}\left(\text{LN}\left(z'_\ell\right)\right) + z'_\ell \tag{6}$$

$$y = \text{LN}\left(\text{z}_L^0\right) \tag{7}$$

where x_{class} is the input PCI data, MLP represents the computational process of the multilayer perceptual machine, and MSA represents the computational process of the multi-head attention mechanism.

It is evident how the single-head attention mechanism is calculated for the MSA computation, as demonstrated in (8).

$$\text{Attention}(Q, K, V) = \text{soft}\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

where the value vector is V, the key vector is K, and the query vector is Q. The single-head attention mechanism extends the computation of the multi-head attention mechanism. The QKV vector matrix is mapped several times by the multi-head attention mechanism, which then combines the vector matrices. The equations (9) and (10) illustrate the computation procedure.

$$\text{MultiHead}(Q, K, V) = \text{Cat}\left(\text{head}_1, \cdots, \text{head}_h\right) \tag{9}$$

$$\text{head}_i = \text{Attention}\left(Q \cdot W_i^Q, K \cdot W_i^K, V \cdot W_i^V\right) \tag{10}$$

To validate the effectiveness of the designed feature extraction network, SectionIV-A demonstrates the structure with experiments.

B. Multi-modal Imaging Feature Fusion

After extracting the features of the three modal, WLI, FI, and PCI data, feature fusion of the three modal features is required. The spatial alignment of the tri-modal data is ensured during acquisition, and the same size is maintained throughout feature extraction and in the resulting feature maps, ensuring that the features from each modality remain aligned. In the feature fusion module, after summing the dimensions of the tri-modal features, only weighting and multiplication operations are applied at corresponding positions, maintaining feature alignment throughout the entire process. Accordingly, we propose the MFF module to perform global and local-level feature fusion on the three modal features and finally output the fused feature $\widehat{f}$, as illustrated in Fig. 2. Initially, an element-wise adding operation along the channel direction is applied to the three retrieved modal characteristics, f_W, f_F, and f_P, as shown in (11).

$$f_{add} = f_W \oplus f_F \oplus f_P \tag{11}$$

Additionally, f_{add} needs to be put into the global local feature fusion (GLFF) module. The GLFF module has two branches: the left branch handles global feature interaction, while the right branch handles local feature interaction. The Sigmoid function is subsequently utilized to output the results after the operation has been added together. Local features can capture specific details and unique information, whereas global features are typically able to capture the general global structure and semantic information of the image. To assist the model in comprehending the overall features, the left branch of the GLFF module implements a global average pooling layer to gather global information about the complete feature map. Furthermore, by using point-wise convolution to interchange and integrate data between various spatial locations or channels, both branches simultaneously alter the dimensionality of the feature maps. In this model, global features f_{add}^{gobal} are extracted through Global Average Pooling, summarizing the overall information from the input data and capturing the global context. On the other hand, local features f_{add}^{local} are obtained using Point-wise Convolution, which retains spatial details and local patterns in the data. The GLFF module integrates both global and local features, leveraging the summarization capability of global features and the detail-preserving nature of local features to enhance feature representation and improve model performance. The GLFF module process is as follows in (12) to (14).

$$f_{add}^{gobal} = BN(PWC(\xi(PWC(GAP(f_{add}))))) \tag{12}$$

$$f_{add}^{local} = BN(PWC(\xi(PWC(f_{add})))) \tag{13}$$

$$f'_{add} = \sigma(f_{add}^{gobal} \oplus f_{add}^{local}) \tag{14}$$

where GAP is global average pooling, PWC is point-wise convolution, ξ is the Rectified Linear Unit (ReLU) activation function, σ is the Sigmoid function, and BN is the Batch Normalization (BN) layer. Here, the input feature tensor shape $C^{'} \times H^{'} \times W^{'}$ is compressed to $\frac{C^{'}}{\gamma} \times 1 \times 1$, then restored to $C^{'} \times H^{'} \times W^{'}$ using point-wise convolution, which exchanges and integrates information.

The totals of the three modal features are subsequently weighed individually using $f_{add}^{'}$, which is the fusion weight information obtained from the GLFF module's output. Finally, the adding operation is carried out to complete the fusion of the three modal features, as in

$$\widehat{f} = (f_{add}^{'} \otimes f_W) \oplus ((1 - f_{add}^{'}) \otimes f_F) \oplus (f_{add}^{'} \otimes f_P). \quad (15)$$

Additionally, to better adapt to various tasks and contexts, the relative prominence of certain modal elements can be dynamically changed. The GLFF output can assist the model in narrowing its focus to the most pertinent data, and by weighting and summing various modal features, the redundancy between features can be minimized. Moreover, the FI has fewer similar features to WLI and PCI, so a one-minus weighting operation is adopted. The MFF enhances the model's capacity for generalization, increases its efficiency, and prevents needless computational and storage overheads.

C. Loss Function

In colorectal cancer patients, metastatic lymph nodes constitute only a small fraction of the total lymph nodes, resulting in a significant imbalance between benign and malignant samples. To address this imbalance, LDAM Loss [28] is employed in this study. Additionally, due to the similarities between benign and malignant lymph nodes observed in intraoperative fluorescence imaging, distinguishing between them can be challenging. To tackle this issue, we use Focal Loss [29]. Finally, we design a novel intra-loss function to enhance sample diversity, further improving model performance by building on the combination of the previous losses. Thus, the new loss function consists of LDAM Loss, Focal Loss, and Intra-Loss, designed to enhance model prediction performance by combining the strengths of these three loss functions.

1) LDAM Loss: The LDAM loss aims to minimize generalization boundaries based on margin. To enhance the generalization performance of the minority class, the classification boundary can be shifted toward the majority class. This loss can be employed with a priori strategies for training with class-imbalanced data, including reweighting or resampling, in place of the traditional cross-entropy function during training [28]. The specific formula is shown in (16).

$$\mathcal{L}_{\text{LDAM}} = -\sum_{i=1}^{N} \log \left(\frac{e^{z_i^y - \Delta_y}}{e^{z_i^y - \Delta_y} + \sum_{j \neq y} e^{z_i^j}} \right) \quad (16)$$

$$where \Delta_j = \frac{\hat{C}}{n_j^{1/4}}, \quad for j \in \{1, \ldots, k\} \quad (17)$$

where $\hat{C}$ is the hyperparameter that needs to be adjusted, n_j is the class j sample size, k is the number of classes and z_i^y is the yth output of the model for the ith sample, which belongs to class y. Moreover, LDAM is reweighted by frequency in order to enhance its performance on class-imbalance learning. In our loss function, the reweighted LDAM is used, as in

$$\mathcal{L}_{\text{LDAM}}^{\text{reweighted}} = -\sum_{i=1}^{N} \omega_i \log \left(\frac{e^{z_i^y - \Delta_y}}{e^{z_i^y - \Delta_y} + \sum_{j \neq y} e^{z_i^j}} \right). \quad (18)$$

2) Focal Loss: Focal loss effectively suppresses the influence of easily classified samples on the loss by reducing their weight and increasing the weight of hard-to-classify samples, thus enhancing the model's focus on difficult-to-classify samples [29]. It addresses class imbalance issues and improves model performance on minority classes. The formula is shown blew

$$\mathcal{L}_{\text{FL}} = -\sum_{i=1}^{N} \omega_i \left(1 - p_i\right)^{\kappa} \log \left(p_i\right) \quad (19)$$

where $(1 - p_i)^{\kappa}$ is a modulating factor, and κ is a hyperparameter needed to be setting. The higher value of $(1 - p_i)^{\kappa}$ means that the model will pay more attention to hard samples with small values of p_i.

3) Intra Loss: In order to bring the intra-class samples closer together and increase the diversity of the samples, we propose the Intra-loss. Initially, the Pearson correlation coefficient (PCC) is employed to determine the similarity between neighboring samples within a class, as shown in

$$\rho_{\mathbf{x}_i \mathbf{x}_{i+1}} = \frac{\text{Cov}\left(f\left(\mathbf{x}_i\right), f\left(\mathbf{x}_{i+1}\right)\right)}{\delta\left(f\left(\mathbf{x}_i\right)\right) \delta\left(f\left(\mathbf{x}_{i+1}\right)\right)} \quad (20)$$

where $f\left(\mathbf{x}_i\right)$ and $f\left(\mathbf{x}_i\right)$ represent the output of the model for samples x_i and x_{i+1}, respectively; $\text{Cov}(\cdot)$ is the covariance function; $\delta(\cdot)$ is the standard deviation function. Therefore, the average similarity of each class defines the feature diversity ϕ_m of class m, as shown blew

$$\rho_m = \frac{\sum_{x_i \in \phi^m} \rho_{x_i x_{i+1}}}{n_m} \quad (21)$$

where n_m is the number of samples belonging to class m, and ϕ^m is the set of sample space on class m. As illustrated in (22), distinct classes are required to have similar data distributions in intra-class losses to concurrently tackle the class imbalance problem of data scarcity and data density.

$$\mathcal{L}_{\text{Intra}} = \sum_{m,k \in \phi \text{ and } m \neq k} \|\rho_m - \rho_k\|_2^2 \quad (22)$$

The final form of the loss function is shown below

$$\mathcal{L}_{\text{All}} = \mathcal{L}_{\text{LDAM}}^{\text{reweighted}} + \alpha \mathcal{L}_{\text{FL}} + \beta \mathcal{L}_{\text{Intra}}. \quad (23)$$

The proposed novel loss function has combined the aforementioned three loss functions, which will address the issues of sample imbalance and differentiation difficulty while also increasing the intra-class variety.

III. EXPERIMENTAL SETTINGS

A. Data

The intraoperative fluorescence dataset employed in this study contains three modalities of WLI, FI, and PCI data, where FI data are NIR-I fluorescence. The lymph node samples used in this study are isolated during colorectal cancer surgeries and imaged intraoperatively. For each lymph node, we capture three types of modality images: white-light imaging (WLI), fluorescence imaging (FI), and pseudo-color imaging (PCI) using the intraoperative fluorescence device (DPM-ENDOSCOPE-3D06 and DPM-0PENCAM-02, Zhuhai Dipu Medical Technology Co., Ltd.). The training and test datasets are provided by Beijing Fengtai Hospital, which supplied multi-modal fluorescence images of isolated lymph nodes from 49 patients. A total of 996 lymph nodes are analyzed, 87 of which are metastatic, while the remaining 909 are non-metastatic, resulting in a negative-to-positive sample ratio of approximately 10:1. In total, 2,988 multi-modal images are collected, with 996 images per modality. The dataset is split into train and test sets in the trials using a 7:3 ratio, and the three modal images adhere to the same dataset division guidelines.

To verify the model's generalization performance and robustness, we acquire multi-modal lymph node images from 19 colorectal cancer patients at the Cancer Hospital, Chinese Academy of Medical Sciences. This external validation set consists of 270 lymph nodes, yielding a total of 810 images, of which 76 nodes are metastatic. Intraoperative isolated lymph nodes are placed on medical placement cloths, in order to reduce the interference of the background on the WLI data, we sequentially perform the background removal operation on all WLI image data via the Photoshop software. To avoid inadvertent removal of the edge information of lymph nodes in the WLI, the background removal operation was not carried out on the PCI data so that the feature extraction network could perform the supplementation corresponding to the missing information during the model training. In the beginning, the image size is adjusted to 224×224, and random horizontal flip and random rotation operations are carried out.

B. Implementation Details

The three modal imaging data sizes input to the model are $(B, 3, 224, 224)$, where B is the batch size. In the feature extraction network, the branching structure for extracting the WLI data is a convolutional layer with an output dimension of 32, immediately followed by 17 inverted residual modules, and finally a convolutional and fully connected layer with an output feature map shape of $(B, 512, 7, 7)$. All the above convolutional layers are immediately followed by the BN layer and the ReLU6 activation function. The branching structure of the extracted FI features uses the network structure of ResNet18 [30] throughout, and the shape of the output feature map is $(B, 512, 7, 7)$. The feature branching structure of the extracted PCI is the ViT structure, which is firstly sliced and mapped to the data, then converted to the form of a long sequence of vectors, immediately followed by stacking of 12 Transformer modules, and finally outputting a feature map with the dimensions $(B, 512, 7, 7)$. Pre-trained weights trained on the ImageNet dataset have been introduced into the complete feature extraction network to speed up model convergence and increase model accuracy.

The three modal feature dimensions $C' \times H' \times W'$ input to the MFF module are $(B, 512, 7, 7)$, followed by element-wise addition, and the output feature dimensions after passing through the MFF module is $(B, 512, 7, 7)$, where the hyperparameter γ in the GLFF module is 8. The final multi-layer perception (MLP) structure of the model is an adaptive pooling layer with an output dimension of $(B, 512, 1, 1)$, followed by two fully connected (FC) layers mapping the feature dimensions from 512 to 2. Moreover, the two FC layers have a ReLU activation function applied and a dropout layer to prevent over-fitting.

The hyperparameters during model training are set as follows: learning rate (lr) is set to 0.001; weight decay (wd) is set to 0.003; batch size (B) is 32; and the dropout layer hyperparameter is set to 0.3. The Adam optimizer is chosen as the optimizer for model training. A total of 100 epochs are set up for the training process. The hyperparameter settings in the loss function align with those in references [28] and [29]. The hyperparameters in the loss function $\mathcal{L}_{\text{All}}$ are set as follows: the weight w_i of $\mathcal{L}_{\text{LDAM}}^{\text{reweighted}}$ and $\mathcal{L}_{\text{FL}}$ are both set to 0.85; $\hat{C}$ in $\mathcal{L}_{\text{LDAM}}^{\text{reweighted}}$ is set to 0.5; and κ in $\mathcal{L}_{\text{FL}}$ is set to 2.0. In Section IV-E, the sensitivity analysis of the hyperparameters indicates that α and β should be set to 1 for optimal model performance. The training process for all experiments is implemented using Pytorch-2.1.0 and is performed on the NVIDIA RTX 4090 GPU with 24 GB of memory.

C. Evaluation Metrics

To validate the performance of our proposed MFI-FFP model for the prediction of colorectal cancer, the main assessment metrics employed are the area under the receiver operating characteristic (ROC) curve (AUC), accuracy (ACC), and F1 score. The ROC curve is a curve with the true positive rate (TPR) as the y-axis and the false positive rate (FPR) as the x-axis. The AUC is the area under the ROC curve, which indicates the model's classification performance under different thresholds. The larger the AUC means the better the model's performance.

IV. RESULTS

A. Comparison of Different Feature Extraction Branches

To evaluate the performance of different network architectures as feature extraction branches for the three modal imaging, we extracted WLI, FI, and PCI data features using the residual block (RB) structure, the inverted residual block (IRB) structure, and the ViT structure, respectively. As shown in Table I, the results of applying the same branching structure to the three modal data are also compared. Consequently, all comparison experiments are evaluated on the test dataset and external validation set across 5 runs. The median of these 5 runs of results is selected for presentation and 95% confidence intervals (CI) are calculated using 1000 trials of the Bootstrap approach.

TABLE I

COMPARISON OF RESULTS OF DIFFERENT FEATURE EXTRACTION BRANCHES FOR EXTRACTING THREE MODAL IMAGING DATA. NUMBERS IN BRACKETS INDICATE 95% CONFIDENCE INTERVALS USING 1000 TRIALS OF THE BOOTSTRAP METHOD

Image mode	+Feature extraction branch		test set			External validation set		
			AUC	ACC	F1	AUC	ACC	F1
WLI+IRB	FI+RB	PCI+ViT	**0.8294** (0.7813,0.8431)	**0.8635** (0.8487,0.8967)	**0.4999** (0.4612, 0.5174)	**0.8437** (0.8006, 0.8769)	0.8111 (0.7630, 0.8356)	**0.6946** (0.6471, 0.7161)
WLI+RB	FI+IRB	PCI+ViT	0.7426 (0.7344,0.7709)	0.7970 (0.7721,0.8219)	0.3093 (0.2733,0.3411)	0.7569 (0.7283,0.7826)	0.7685 (0.7215, 0.7852)	0.5323 (0.5067,0.5562)
WLI+ViT	FI+RB	PCI+IRB	0.8054 (0.7935,0.8283)	0.8672 (0.8462,0.8882)	0.4236 (0.3920,0.4542)	0.8149 (0.7636, 0.8312)	**0.8296** (0.7407, 0.8407)	0.6290 (0.5928, 0.6548)
WLI+ViT	FI+IRB	PCI+RB	0.7691 (0.7517,0.8036)	0.8192 (0.7953,0.8431)	0.3544 (0.3271,0.3815)	0.7023 (0.6625, 0.7224)	0.7815 (0.7626, 0.8163)	0.5203 (0.4916, 0.5526)
WLI+IRB	FI+ViT	PCI+RB	0.7600 (0.7460,0.7915)	0.7934 (0.7683,0.8185)	0.3467 (0.3149,0.3804)	0.7792 (0.7582, 0.8180)	0.7556 (0.7237, 0.7837)	0.6118 (0.5974, 0.6402)
WLI+RB	FI+ViT	PCI+IRB	0.7853 (0.7510,0.8161)	0.8229 (0.7992,0.8466)	0.3688 (0.3422,0.4013)	0.7424 (0.7135,0.7765)	0.7481 (0.7167, 0.7704)	0.5211 (0.4902, 0.5457)
WLI+IRB	FI+ IRB	PCI+IRB	0.7089 (0.6704,0.7252)	0.7528 (0.7261,0.7795)	0.2708 (0.2437,0.3054)	0.7526 (0.7146, 0.7856)	0.7556 (0.7237, 0.7874)	0.5238 (0.4988, 0.5565)
WLI+RB	FI+RB	PCI+RB	0.6268 (0.6011,0.6544)	0.6753 (0.6463,0.7043)	0.2317 (0.2048,0.2573)	0.6681 (0.6402, 0.6912)	0.6778 (0.6470, 0.7031)	0.4710 (0.4416, 0.5189)
WLI+ViT	FI+ViT	PCI+ViT	0.6357 (0.6002,0.6603)	0.6827 (0.6539,0.7115)	0.2275 (0.1983,0.2592)	0.6719 (0.6389, 0.7007)	0.6926 (0.6741, 0.7278)	0.4645 (0.4327,0.4958)

The comparison results of the three modal imaging data extracted by different feature extraction branches show that extracting WLI data features using IRB structure, FI data features using RB structure, and PCI data features using ViT structure demonstrated the best performance, with the AUC of 0.8294 (95% CI 0.7813-0.8431), the ACC of 0.8669 (95% CI 0.8487-0.8967), and the F1 score of 0.4999 (95% CI 0.4612-0.5174) on the test set. Similarly, the model demonstrates strong performance on the external validation set, further confirming its generalization capability and robustness. Furthermore, there is relatively high efficiency in the experimental results when exchanging the WLI and PCI data feature extraction branches, which achieves the AUC of 0.8054 (95% CI 0.7935-0.8283), the ACC of 0.8672 (95% CI 0.8462-0.8882), and the F1 score of 0.4236 (95% CI 0.3920-0.4542) on the test set. The ACC on the external validation set reaches 0.8296. The structure of the PCI data is similar to that of the WLI data, but it is mixed with some background noise and lacks some lymph node texture information after fluorescence information mapping. As a result, the global features cannot be effectively extracted using the IRB structure, which slightly reduces the overall performance of the model. Additionally, the last three rows of Table I demonstrate that even though similar features can be extracted using the same feature extraction structure in all three modalities, the unique features of each modality cannot be adequately represented.

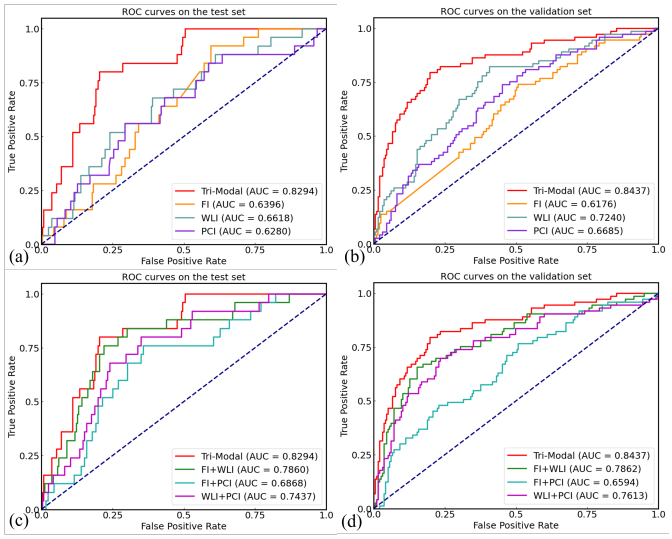

Fig. 3. The ROC curves for comparison results of uni-modal and multi-modal images. (a) The ROC curves for the results of the comparison between the tri-modal and uni-modal data on the test set.(b) The ROC curves for the results of the comparison between the tri-modal and uni-modal data on the validation set. (c) The ROC curves for the results of the comparison between the tri-modal and bi-modal data on the test set. (d) The ROC curves for the results of the comparison between the tri-modal and bi-modal data on the validation set.

TABLE II
COMPARISON RESULTS OF SINGLE-MODAL, MULTI-MODAL IMAGING. NUMBERS IN BRACKETS INDICATE 95% CONFIDENCE INTERVALS USING 1000 TRIALS OF THE BOOTSTRAP METHOD

Image mode	Params(M)	FLOPs(G)	test set			External validation set		
			AUC	ACC	F1	AUC	ACC	F1
Tri-modal	122.40	13.47	**0.8294** (0.7813,0.8431)	**0.8635** (0.8487,0.8967)	**0.4999** (0.4612, 0.5174)	**0.8437** (0.8006, 0.8769)	**0.8111** (0.7630, 0.8356)	**0.6946** (0.6471, 0.7161)
FI+WLI	45.80	2.19	0.7860 (0.7678,0.8174)	0.8192 (0.8087,0.8561)	0.3725 (0.3441,0.3982)	0.7862 (0.7501, 0.8218)	0.7889 (0.7407, 0.8133)	0.5512 (0.5244, 0.5952)
FI+PCI	88.03	13.13	0.6868 (0.6613,0.7175)	0.7726 (0.7365,0.7978)	0.2568 (0.2237,0.2854)	0.6594 (0.6271,0.6830)	0.6778 (0.6410, 0.7034)	0.4459 (0.4103, 0.4771)
WLI+PCI	111.22	11.67	0.7437 (0.7228,0.7725)	0.8376 (0.8007,0.8524)	0.3364 (0.3117,0.3753)	0.7613 (0.7407, 0.7935)	0.7815 (0.7541, 0.8178)	0.5426 (0.5147,0.5702)
FI	11.31	1.82	0.6396 (0.6194,0.6545)	0.7653 (0.7401,0.7942)	0.2049 (0.1865,0.2407)	0.6176 (0.5845,0.6411)	0.6110 (0.5762,0.6357)	0.3721 (0.3378,0.3941)
WLI	3.01	0.33	0.6618 (0.6507,0.6772)	0.7601 (0.7232,0.7897)	0.2574 (0.2235,0.2823)	0.7240 (0.6949,0.7526)	0.6926 (0.6724,0.7276)	0.5030 (0.4786,0.5317)
PCI	57.82	11.29	0.6280 (0.6057,0.6419)	0.7437 (0.7112,0.7617)	0.2016 (0.1783,0.2335)	0.6352,0.6974)	0.6352,0.6974)	0.4177 (0.3747,0.4435)

B. Comparison of Single-Modal, Multi-Modal Imaging

To assess the effectiveness of fusing feature information from tri-modal imaging compared to fusing feature information from bi-modal or even uni-modal imaging, some comparative experiments using only bi-modal or uni-modal imaging are conducted in this study. As shown in Table II, the MFI-FFP model can more comprehensively learn the different feature information of the current corresponding lymph node by combining the feature information from the three modal imaging, WLI, FI, and PCI, resulting in a more effective way to distinguish LNM. The experimental results demonstrate that the network learns best and the model performs optimally when the three modal imaging data serve as inputs, obtaining the ACC of 0.8782 (95% CI 0.8487-0.8967), the AUC of 0.8294 (95% CI 0.7813-0.8431), and the F1 score of 0.4999 (95% CI 0.4612-0.5174) on the test set. In the bi-modal imaging data feature fusion experiments, the model effect of combining WLI and FI demonstrates better performance, achieving an AUC of 0.7860 (95% CI 0.7678-0.8174) and the ACC of 0.8192 (95% CI 0.8087-0.8561) on the test set, and an AUC of 0.7862 (95% CI 0.7501-0.8218) and the ACC of 0.7889 (95% CI 0.7407-0.8133) on the validation set, which indicates that the WLI and FI data contain minimal feature overlap and possess distinct information. Furthermore, the results from the unimodal imaging comparison reveal that the model trained on WLI data alone performed significantly better than those trained on the other two modalities. This not only demonstrates that the WLI data contains a substantial amount of valid lymph node feature information but also indirectly highlights the importance of background removal.

Fig.3 illustrates the ROC curves comparing the tri-modal imaging with the bi-modal and uni-modal imaging experiments on both the test set and validation set. It can be concluded that the classification performance of the tri-modal model is the most superior, and the uni-modal model performs the weakest. The results demonstrate that the characteristics of lymph nodes can be captured as comprehensively as possible using the information from the tri-modal imaging data, which

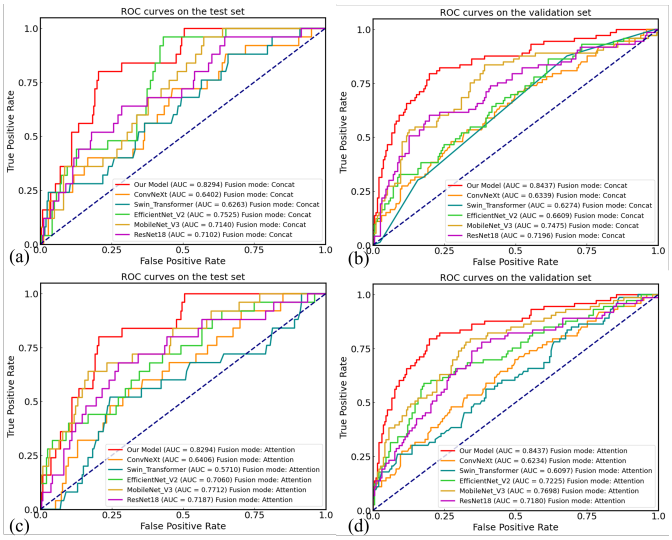

Fig. 4. The ROC curves of the results of our model compared to other classification models. (a) The ROC curves of the results of our model compared to other classification models with features concatenated in dimension on the test set. (b) The ROC curves of the results of our model compared to other classification models with features concatenated in dimension on the validation set. (c) The ROC curves of the results of our model compared to other classification models with attention-based features fusion method on the test set. (d) The ROC curves of the results of our model compared to other classification models with attention-based features fusion method on the validation set.

is much more effective than the bi-modal and uni-modal imaging data and directly enhances the model performance and accuracy.

C. Comparison with Various Classification Models

Comparison of other superior classification models with our proposed MFI-FFP model on the task of intraoperative fluorescence LNM prediction in CRC can visually and efficiently illustrate the excellent performance of our model. Since there are no multi-modal models that specifically deal with intraoperative fluorescence data from lymph nodes in CRC, we choose some image classification models that exhibit superior performance on natural images as comparison models. Several outstanding research works have been conducted on natural image classification tasks; among these, the Resnet18 [30], the Swin-Transformer [31], the ConvNeXt [32], the Efficientnet-v2 [33], and the Mobilenet-v3 [34] are selected as comparison models.

All comparative models are pre-trained on the ImageNet dataset to boost the convergence of these models. The three modal features produced by the comparative models are fused for each of the two fusion methods, one for concatenating in the dimensional direction and one for attention-based fusion (WLI features serve as the query tensors, FI features as the key tensors, and PCI features as the value tensors). The experiment compares these two common fusion methods to prove our model's performance. The loss functions for all comparison models training are all set to a weighted cross-entropy loss function with weights of 0.85. The five comparison models include both the traditional convolutional network structure and the Transformer structure, which has been quite remarkable in recent years, and the final comparison results are shown in Tables III and IV. The results indicate that the effect of our model works best on the intraoperative fluorescence tri-modal imaging dataset, followed by the training results of the Mobilenet-v3 model with attention-based fusion, which also demonstrated good results, achieving an AUC of 0.7712 (95% CI 0.7369-0.7915) and an ACC of 0.8266 (95% CI 0.8038-0.8559) on the test set, and an AUC of 0.7698 (95% CI 0.7257-0.7988) and an ACC of 0.7630 (95% CI 0.7315-0.7953) on the validation set. The floating point operations per second (FLOPs) of our model consume less computational resources compared to the pure Transformer structure model, but this is slightly worse than the convolutional network structure model, which is caused by the fact that the Transformer structure with a large number of parameters is also included in our model. Fig. 4 shows the ROC curves comparing the results of multiple models using different fusion methods, visually illustrating the superior performance of our model. All the above comparison experiments are trained on the dataset for 5 runs to avoid serendipity in the experimental results.

D. Ablation Experiments

In addition to conducting comparative experiments to validate the efficient performance of our model, we also design three ablation studies to evaluate the performance impact of the innovative part on the MFI-FFP model. The first ablation study is performed on each part of the losses in the designed loss function to verify the impact of each of these loss parts on the model training process, which is based on the presence of the fusion module MFF. In addition, the second ablation study is performed on the designed loss function and the fusion module MFF to verify the impact of each component on our model. The final ablation experiment demonstrates the effectiveness of different components by removing the feature extraction branch and the MFF module, confirming the significance of each in enhancing the model's overall performance.

1) Ablation Study for Analyzing the Loss Function: Table V presents the general results of the ablation experiment conducted on each of the three loss parts, $\mathcal{L}_{\mathrm{LDAM}}^{\mathrm{reweighted}}$, $\mathcal{L}_{\mathrm{FL}}$, and $\mathcal{L}_{\mathrm{Intra}}$ in the first ablation study. The results indicated that all three loss components of the loss function $\mathcal{L}_{\mathrm{All}}$ applied during model training performed the best, achieving the highest AUC and ACC. The ablation experiment with $\mathcal{L}_{\mathrm{Intra}}$ alone as the entire loss function performs the worst, this is because the loss is not responsible for solving the sample imbalance problem and the problem of difficulty in distinguishing between samples. Applying only the $\mathcal{L}_{\mathrm{LDAM}}^{\mathrm{reweighted}}$ and $\mathcal{L}_{\mathrm{FL}}$ losses can reach an AUC of 0.7576 (95% CI 0.7211-0.7841) on the test set and an AUC of 0.8005 (95% CI 0.7707-0.8373) on the validation set, and the result after adding the $\mathcal{L}_{\mathrm{Intra}}$ loss shows that the $\mathcal{L}_{\mathrm{Intra}}$ loss is effective in improving the diversity of the samples, which indirectly enhances the overall model effect. The sample imbalance issue in the validation set is not as pronounced as in the test set, and the model's performance on the validation set is superior to that on the test set. This indirectly highlights the effectiveness of the designed loss function in addressing sample imbalance. Additionally, the impact increase is more noticeable when a single $\mathcal{L}_{\mathrm{LDAM}}^{\mathrm{reweighted}}$ or $\mathcal{L}_{\mathrm{FL}}$ is combined with $\mathcal{L}_{\mathrm{Intra}}$ loss, respectively.

2) Ablation Study for the Loss Function and the MFF module: Table VI demonstrates the results of the second ablation study, where the model that uses neither the newly designed loss function (applying a cross-entropy loss function with a weight of 0.85) nor the MFF has the worst performance, with the AUC of only 0.6933 (95% CI 0.6503-0.7241) on the test set and the AUC of 0.6518 (95% CI 0.6163-0.6789) on the validation set. Moreover, it is evident from the experimental results that the model has a significant performance improvement effect by applying the new loss function alone and the MFF fusion module alone. The simultaneous usage of the newly designed loss function and the fusion module MFF enables the model to reach its highest performance.

3) Ablation Study for the feature extraction branches and the MFF module: Table VII presents the results of the final ablation experiment, showing that the worst performance is observed when the MFF module is removed, while the three feature extraction branches maintain the same network structure. Once the MFF module was added, all models experienced a performance improvement, demonstrating its effectiveness. Furthermore, when our custom-designed feature extraction branches are used as the backbone network, the overall results surpass those of other branch designs. This validates the design of the inverse residual network for extracting WLI features, the residual network for FI features, and the ViT network for PCI

TABLE III
COMPARISON RESULTS OF VARIOUS CLASSIFICATION MODELS WITH OUR MODEL ON THE TEST SET.NUMBERS IN BRACKETS INDICATE 95% CONFIDENCE INTERVALS USING 1000 TRIALS OF THE BOOTSTRAP METHOD

Model	Params(M)	FLOPs(G)	Fusion mode					
			Concat			Attention		
			AUC	ACC	F1	AUC	ACC	F1
Swin-Transformer	103.08	17.28	0.6263 (0.5997,0.6672)	0.6826 (0.6384,0.6824)	0.2185 (0.1831,0.2476)	0.5710 (0.5576,0.6173)	0.6273 (0.5973,0.6415)	0.1985 (0.1744,0.2223)
ConvNeXt	152.25	26.09	0.6402 (0.6250,0.6654)	0.7159 (0.6974,0.7306)	0.2020 (0.1797,0.2351)	0.6406 (0.6023,0.6829)	0.6900 (0.6563,0.7127)	0.2174 (0.1812,0.2447)
Efficientnet v2	65.25	8.70	0.7525 (0.7231,0.7876)	0.8118 (0.7823,0.8450)	0.3103 (0.2811,0.3379)	0.7060 (0.6767,0.7426)	0.7232 (0.6908,0.7541)	0.3200 (0.2942,0.3518)
Mobilenet v3	16.44	0.23	0.7140 (0.6763,0.7318)	0.7675 (0.7454,0.7970)	0.3214 (0.2957,0.3528)	0.7712 (0.7369,0.7915)	0.8266 (0.8038,0.8559)	0.3434 (0.3120,0.3788)
Resnet18	36.3	5.47	0.7102 (0.6714,0.7418)	0.7860 (0.7601,0.8007)	0.2985 (0.2714,0.3211)	0.7187 (0.6860,0.7412)	0.7380 (0.7066,0.7712)	0.3063 (0.2714,0.3347)
Our model	122.40	13.47	AUC:**0.8294** (0.7813,0.8431)		ACC:**0.8635** (0.8487,0.8967)		F1:**0.4999** (0.4612, 0.5174)	

TABLE IV
COMPARISON RESULTS OF VARIOUS CLASSIFICATION MODELS WITH OUR MODEL ON THE EXTERNAL VALIDATION SET.NUMBERS IN BRACKETS INDICATE 95% CONFIDENCE INTERVALS USING 1000 TRIALS OF THE BOOTSTRAP METHOD

Model	Params(M)	FLOPs(G)	Fusion mode					
			Concat			Attention		
			AUC	ACC	F1	AUC	ACC	F1
Swin-Transformer	103.08	17.28	0.6274 (0.5916, 0.6531)	0.6444 (0.5941, 0.6578)	0.4070 (0.3748, 0.4317)	0.6097 (0.5842,0.6593)	0.6259 (0.6078, 0.6415)	0.4094 (0.3877, 0.4383)
ConvNeXt	152.25	26.09	0.6339 (0.6141,0.6754)	0.6481 (0.5978, 0.6615)	0.4379 (0.3971, 0.4532)	0.6234 (0.5719, 0.6314)	0.6407 (0.6154, 0.6723)	0.4393 (0.4011, 0.4607)
Efficientnet v2	65.25	8.70	0.6609 (0.6303,0.6927)	0.6852 (0.6448, 0.7012)	0.4371 (0.3704, 0.4662)	0.7225 (0.6957, 0.7544)	0.7363 (0.7003, 0.7658)	0.5500 (0.5302, 0.5953)
Mobilenet v3	16.44	0.23	0.7475 (0.7232,0.7853)	0.7815 (0.7412, 0.8106)	0.5693 (0.6471, 0.7161)	0.7698 (0.7257, 0.7988)	0.7630 (0.7315, 0.7953)	0.5294 (0.3996, 0.5415)
Resnet18	36.3	5.47	0.7196 (0.6878,0.7421)	0.7370 (0.7141, 0.7742)	0.5103 (0.4917, 0.5564)	0.7180 (0.6867,0.7472)	0.7037 (0.6807, 0.7441)	0.5294 (0.4853, 0.5411)
Our model	122.40	13.47	AUC: **0.8437** (0.8006, 0.8769)		ACC: **0.8111** (0.7630, 0.8356)		F1: **0.6946** (0.6471, 0.7161)	

features.

E. Hyperparameter sensitivity analysis

To verify the impact of the loss function hyperparameters on model performance, we set the values of α and β within the range [0.01, 0.05, 0.1, 0.5, 1.0] for hyperparameter sensitivity analysis. As shown in Fig. 6, small values for α and β negatively affect overall model performance, resulting in an AUC of only about 0.75 on the test set, with similar poor performance on the validation set. In contrast, the model performs better when α and β are close to 1, indicating improved results when the weights of the three individual losses in the total loss function are similar. This demonstrates that the model is sensitive to hyperparameters α and β, achieving optimal performance when both are set to 1.

F. Comparison with clinical experts

To assess the clinical value of the model, we invited five clinicians with experience in colorectal cancer treatment to evaluate lymph node metastasis based on multi-modal fluorescence images from the external validation set. This group included experts 1 and 4, who have less than 10 years of clini-

TABLE V

ABLATION EXPERIMENTS FOR EACH LOSS COMPONENT IN THE LOSS FUNCTION. NUMBERS IN BRACKETS INDICATE 95% CONFIDENCE INTERVALS USING 1000 TRIALS OF THE BOOTSTRAP METHOD

$\mathcal{L}_{\text{LDAM}}^{\text{reweighted}}$	$\mathcal{L}_{\text{FL}}$	$\mathcal{L}_{\text{Intra}}$	test set			External validation set		
			AUC	ACC	F1	AUC	ACC	F1
✓			0.7332 (0.7208,0.7463)	0.8044 (0.7897,0.8303)	0.3357 (0.3172,0.3617)	0.7744 (0.7481, 0.8084)	0.7370 (0.6989, 0.7526)	0.6120 (0.6006, 0.6587)
	✓		0.6452 (0.6390,0.6777)	0.7491 (0.7196,0.7749)	0.2548 (0.2214,0.2758)	0.6565 (0.6171,0.6833)	0.7037 (0.6704, 0.7215)	0.4521 (0.4278, 0.4713)
		✓	0.6238 (0.6143,0.6378)	0.7122 (0.1697,0.7417)	0.2015 (0.1829,0.2342)	0.5750 (0.5478,0.6063)	0.5519 (0.5204, 0.5852)	0.4265 (0.3947, 0.4512)
✓	✓		0.7576 (0.7211,0.7841)	0.7971 (0.7749,0.8450)	0.4020 (0.3787,0.4311)	0.8005 (0.7707,0.8373)	0.8011 (0.7754, 0.8215)	0.6250 (0.6010, 0.6567)
✓		✓	0.7403 (0.7278,0.7620)	0.8155 (0.7712,0.8229)	0.3664 (0.3305,0.3875)	0.7837 (0.7506,0.8173)	0.7926 (0.7663, 0.8211)	0.6216 (0.5851, 0.6439)
	✓	✓	0.6816 (0.6267,0.7059)	0.7528 (0.7269,0.7934)	0.3043 (0.28059,0.3274)	0.7276 (0.6808, 0.7503)	0.7778 (0.7415, 0.7952))	0.5714 (0.5317, 0.5908)
✓	✓	✓	**0.8294** (0.7813,0.8431)	**0.8635** (0.8487,0.8967)	**0.4999** (0.4612, 0.5174)	**0.8437** (0.8006, 0.8769)	**0.8111** (0.7630, 0.8356)	**0.6946** (0.6471, 0.7161)

TABLE VI

ABLATION EXPERIMENTS WITH LOSS FUNCTION AND FUSION MODULE MFF. NUMBERS IN BRACKETS INDICATE 95% CONFIDENCE INTERVALS USING 1000 TRIALS OF THE BOOTSTRAP METHOD

$\mathcal{L}_{\text{All}}$	MFF	test set			External validation set		
		AUC	ACC	F1	AUC	ACC	F1
		0.6933 (0.6503,0.7241)	0.7380 (0.7085,0.7638)	0.3112 (0.2816,0.3384)	0.6518 (0.6163,0.6789)	0.6741 (0.6578, 0.7015)	0.4430 (0.4208,0.4745)
	✓	0.7539 (0.7272,0.7806)	0.8087 (0.7860,0.8376)	0.4153 (0.3847,0.4402)	0.7552 (0.7228, 0.7893)	0.7556 (0.7315, 0.7863)	0.5714 (0.5536, 0.6082)
✓		0.7472 (0.7284,0.7763)	0.7934 (0.7712,0.8192)	0.4094 (0.3882,0.4245)	0.7734 (0.7407,0.7962)	0.7852 (0.7641, 0.8015)	0.6329 (0.6102, 0.6647)
✓	✓	**0.8294** (0.7813,0.8431)	**0.8635** (0.8487,0.8967)	**0.4999** (0.4612, 0.5174)	**0.8437** (0.8006, 0.8769)	**0.8111** (0.7630, 0.8356)	**0.6946** (0.6471, 0.7161)

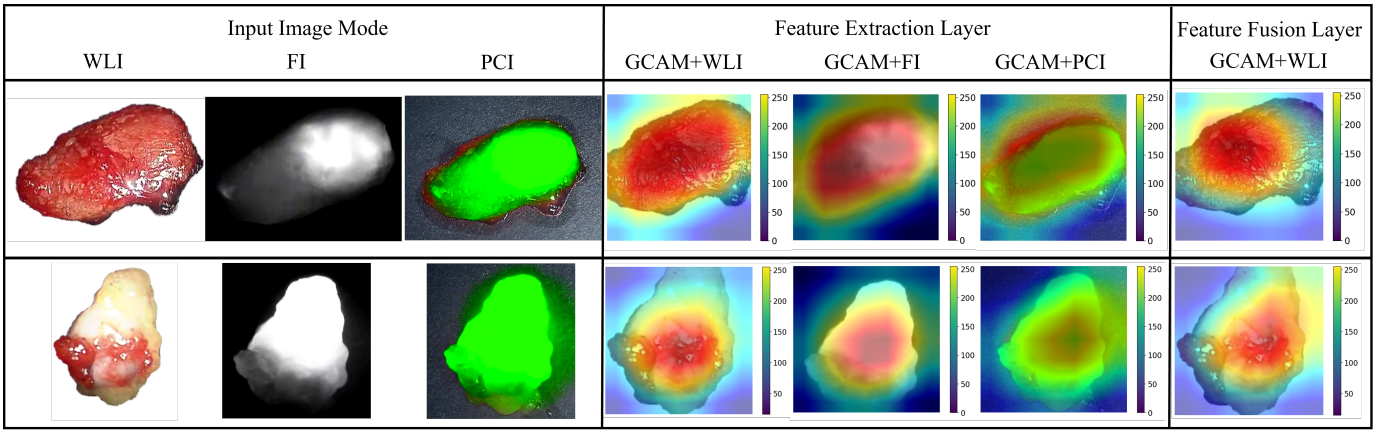

Fig. 5. The schematic GCAM heatmap of lymph node samples. The first row shows the three modal images of the sample without lymph node metastasis, as well as the GCAM mapping heatmaps of the three modalities of the feature extraction layer and the GCAM heatmap mapping of the WLI of the fusion layer. The second row shows the three modality images of the sample where lymph node metastasis has occurred, as well as the GCAM mapping heatmap of the three modalities of the feature extraction layer and the GCAM heatmap mapping of the WLI of the fusion layer.

TABLE VII
ABLATION EXPERIMENTS WITH FEATURE EXTRACTION BRANCHES AND FUSION MODULE MFF. NUMBERS IN BRACKETS INDICATE 95% CONFIDENCE INTERVALS USING 1000 TRIALS OF THE BOOTSTRAP METHOD

Feature extraction branches	MFF	test set			External validation set		
		AUC	ACC	F1	AUC	ACC	F1
Tri-IRB		0.6810 (0.6553, 0.7067)	0.6974 (0.6785, 0.7233)	0.2407 (0.2125, 0.2732)	0.7013 (0.6706, 0.7218)	0.7037 (0.6842, 0.7221)	0.4407 (0.4385, 0.4704)
Tri-RB		0.6154 (0.5828, 0.6331)	0.6494 (0.6234,0.6615)	0.2180 (0.2025, 0.2323)	0.6281 (0.5979,0.6358)	0.6481 (0.6259,0.6728)	0.4379 (0.4124,0.4502)
Tri-ViT		0.6193 (0.6021,0.6454)	0.6531 (0.6215, 0.6752)	0.2074 (0.1895,0.2214)	0.6551 (0.6358,0.6765)	0.6852 (0.6549,0.7036)	0.4056 (0.2819,0.4276)
Tri-IRB	✓	0.7089 (0.6704,0.7252)	0.7528 (0.7261,0.7795)	0.2708 (0.2437,0.3054)	0.7526 (0.7146, 0.7856)	0.7556 (0.7237, 0.7874)	0.5238 (0.4988, 0.5565)
Tri-RB	✓	0.6268 (0.6011,0.6544)	0.6753 (0.6463,0.7043)	0.2317 (0.2048,0.2573)	0.6681 (0.6402, 0.6912)	0.6778 (0.6470, 0.7031)	0.4710 (0.4416, 0.5189)
Tri-ViT	✓	0.6357 (0.6002,0.6603)	0.6827 (0.6539,0.7115)	0.2275 (0.1983,0.2592)	0.6719 (0.6389, 0.7007)	0.6926 (0.6741, 0.7278)	0.4645 (0.4327,0.4958)
IRB_RB_ViT		0.7472 (0.7284,0.7763)	0.7934 (0.7712,0.8192)	0.4094 (0.3882,0.4245)	0.7734 (0.7407,0.7962)	0.7852 (0.7641, 0.8015)	0.6329 (0.6102, 0.6647)
IRB_RB_ViT	✓	**0.8294** (0.7813,0.8431)	**0.8635** (0.8487,0.8967)	**0.4999** (0.4612, 0.5174)	**0.8437** (0.8006, 0.8769)	**0.8111** (0.7630, 0.8356)	**0.6946** (0.6471, 0.7161)

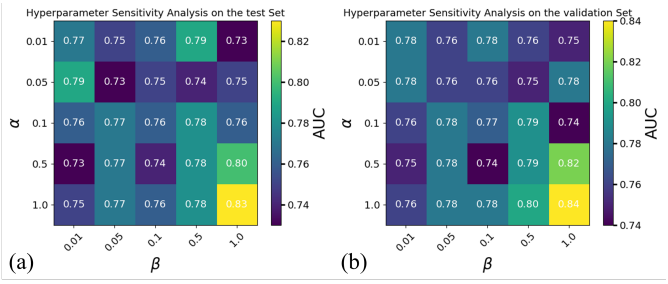

Fig. 6. Hyperparameter sensitivity analysis. (a) Results of the sensitivity analysis of hyperparameters α and β in the loss function on the test set. (b) Results of the sensitivity analysis of hyperparameters α and β in the loss function on the validation set.

TABLE VIII
COMPARISON RESULTS OF CLINICAL EXPERTS WITH OUR MODEL ON THE EXTERNAL VALIDATION SET

Clinical Expert	SEN	SPE	ACC	PREC	F1
Expert 1	0.1447	**0.9381**	0.7148	0.4783	0.2222
Expert 2	0.4342	0.9021	0.7704	0.6346	0.5156
Expert 3	0.4211	0.8247	0.7111	0.4848	0.4507
Expert 4	0.2632	0.7371	0.6037	0.2817	0.2721
Expert 5	0.3947	0.8557	0.7259	0.5172	0.4478
Our model	**0.7632**	0.8299	**0.8111**	**0.6374**	**0.6946**

cal experience, and experts 2, 3, and 5, who have more than 10 years of clinical experience. The clinicians provided a binary judgment on whether a lymph node is metastatic based on its tri-modal imaging. The evaluation metrics include sensitivity(SEN), specificity(SPE), accuracy(ACC), precision(PREC), and F1 score. As shown in Table VIII, the results demonstrate that our model outperforms the clinicians in sensitivity, accuracy, precision, and F1 score. Notably, expert 1 achieved a specificity of 0.9381, but a sensitivity of only 0.1447, which is attributed to his tendency to classify most lymph nodes as non-metastatic. Overall, the results demonstrate that our model is more effective at detecting potential features in images compared to clinical experts, making it a more accurate and efficient approach than traditional imaging-based clinical judgment.

G. Interpretative Visualization

To increase the interpretability of our model network, we utilize the Gradient-weighted class activation mapping (GCAM) method [35], known as a technology for the visualization and interpretation of deep learning models, to increase the understanding of the feature extraction and feature fusion layers of the model. Fig. 5 displays two samples: one with LNM and the other without. The first three columns display the three modal images of the lymph node samples. Columns 4 to 6 display the results of GCAM heat map mapping when the three modal imaging features are extracted on each of the three feature extraction layers, and the final column reveals the GCAM heat map mapped image on the WLI image on the fusion layer.

The results reveal that the feature extraction branch of the convolutional network structure pays more attention to information such as the surface texture of the image, while the feature extraction branch of the Transformer structure pays more attention to the global location information supplementing the lymph node edge feature information. Furthermore, the fused features are combined with modifications to the region of interest by the feature fusion layer, which highlights and concretizes the characteristics in crucial places. For images of lymph nodes that have metastasized, our model pays more attention to the weird irregular texture and shape features in the samples.

V. DISCUSSION AND CONCLUSION

In this study, we propose the MFI-FFP model for LNM prediction in CRC using intraoperative fluorescence multi-modal imaging. Our model first extracts three modal data features, followed by the fusion of the extracted three modal features, and eventually performs LNM in CRC prediction based on the fused features. The suggested MFI-FFP model exhibited promise for real-time LNM diagnosis in the future by performing satisfactorily in the task of predicting LNM in CRC using intraoperative fluorescence multi-model imaging.

In the feature extraction part of our model, the corresponding special extraction branches are set according to the characteristics of the three modal data, respectively. Furthermore, experimental results indicate it is frequently more effective in building distinct feature extraction structures based on various modal imaging feature information than in using the same structure. The three modal features extracted in the previous stage are features fused in the MFF module in the model, which incorporates global and local features to boost model performance. Furthermore, there is a significant sample imbalance issue because the majority of the lymph nodes in the data sample are benign, and relatively few of them are malignant. According to the above problems, we design a novel loss function, in which the LDAM loss alleviates the sample imbalance problem, the focal loss reduces the difficulty in distinguishing the samples, and the proposed intra-loss increases the diversity of the samples and thus improves the performance of the model.

To predict LNM in CRC, our work leverages intraoperative fluorescence multi-modal imaging. By combining three modal imaging characteristics, the model can effectively learn lymph node feature information and exhibit outstanding prediction ability. Since little research has combined deep learning with intraoperative FI for identifying LNM, utilizing deep learning to analyze intraoperative FI for various downstream tasks has far-reaching implications and potential in the medical field.

REFERENCES

[1] R. L. Siegel, N. S. Wagle, A. Cercek, R. A. Smith, and A. Jemal, "Colorectal cancer statistics, 2023," *CA: a cancer journal for clinicians*, vol. 73, no. 3, pp. 233–254, 2023.

[2] E. Dekker, P. Tanis, J. Vleugels, P. Kasi, M. Wallace, and G. Dekker, "Colorectal cancer," *The Lancet*, vol. 394, pp. 1467–1480, 2019.

[3] X. Guan, S. Jiao, R. Wen, G. Yu, J. Liu, D. Miao, R. Wei, W. Zhang, L. Hao, L. Zhou *et al.*, "Optimal examined lymph node number for accurate staging and long-term survival in rectal cancer: a population-based study," *International Journal of Surgery*, vol. 109, no. 8, pp. 2241–2248, 2023.

[4] G. J. Chang, M. A. Rodriguez-Bigas, C. Eng, and J. M. Skibber, "Lymph node status after neoadjuvant radiotherapy for rectal cancer is a biologic predictor of outcome," *Cancer*, vol. 115, no. 23, pp. 5432–5440, 2009.

[5] Y. Eroğlu, M. Yildirim, and A. Çinar, "Convolutional neural networks based classification of breast ultrasonography images by hybrid method with respect to benign, malignant, and normal using mrmr," *Computers in biology and medicine*, vol. 133, p. 104407, 2021.

[6] J. Li, D. Dong, M. Fang, R. Wang, J. Tian, H. Li, and J. Gao, "Dual-energy ct–based deep learning radiomics can improve lymph node metastasis risk prediction for gastric cancer," *European Radiology*, vol. 30, pp. 2324–2333, 2020.

[7] D. Dong, M.-J. Fang, L. Tang, X.-H. Shan, J.-B. Gao, F. Giganti, R.-P. Wang, X. Chen, X.-X. Wang, D. Palumbo *et al.*, "Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study," *Annals of oncology*, vol. 31, no. 7, pp. 912–920, 2020.

[8] M. Ahmed, A. D. Purushotham, and M. Douek, "Novel techniques for sentinel lymph node biopsy in breast cancer: a systematic review," *The lancet oncology*, vol. 15, no. 8, pp. e351–e362, 2014.

[9] J. Zhao, H. Wang, Y. Zhang, R. Wang, Q. Liu, J. Li, X. Li, H. Huang, J. Zhang, Z. Zeng *et al.*, "Deep learning radiomics model related with genomics phenotypes for lymph node metastasis prediction in colorectal cancer," *Radiotherapy and Oncology*, vol. 167, pp. 195–202, 2022.

[10] L. Kiehl, S. Kuntz, J. Höhn, T. Jutzi, E. Krieghoff-Henning, J. N. Kather, T. Holland-Letz, A. Kopp-Schneider, J. Chang-Claude, A. Brobeil *et al.*, "Deep learning can predict lymph node status directly from histology in colorectal cancer," *European Journal of Cancer*, vol. 157, pp. 464–473, 2021.

[11] J. Li, Y. Zhou, P. Wang, H. Zhao, X. Wang, N. Tang, and K. Luan, "Deep transfer learning based on magnetic resonance imaging can improve the diagnosis of lymph node metastasis in patients with rectal cancer," *Quantitative Imaging in Medicine and Surgery*, vol. 11, no. 6, p. 2477, 2021.

[12] C. Chi, Y. Du, J. Ye, D. Kou, J. Qiu, J. Wang, J. Tian, and X. Chen, "Intraoperative imaging-guided cancer surgery: from current fluorescence molecular imaging methods to future multi-modality imaging technology," *Theranostics*, vol. 4, no. 11, p. 1072, 2014.

[13] K. Wang, Y. Du, Z. Zhang, K. He, Z. Cheng, L. Yin, D. Dong, C. Li, W. Li, Z. Hu *et al.*, "Fluorescence image-guided tumour surgery," *Nature Reviews Bioengineering*, vol. 1, no. 3, pp. 161–179, 2023.

[14] T. Burghgraef, A. Zweep, D. Sikkenk, M. van der Pas, P. Verheijen, and E. Consten, "In vivo sentinel lymph node identification using fluorescent tracer imaging in colon cancer: a systematic review and meta-analysis," *Critical Reviews in Oncology/Hematology*, vol. 158, p. 103149, 2021.

[15] S. H. Emile, H. Elfeki, M. Shalaby, A. Sakr, P. Sileri, S. Laurberg, and S. D. Wexner, "Sensitivity and specificity of indocyanine green near-infrared fluorescence imaging in detection of metastatic lymph nodes in colorectal cancer: Systematic review and meta-analysis," *Journal of Surgical Oncology*, vol. 116, no. 6, pp. 730–740, 2017.

[16] S. R. Thammineedi, A. R. Saksena, S. Nusrath, R. R. Iyer, S. Shukla, S. C. Patnaik, R. P. Reddy, N. Boleneni, R. M. Sharma, L. Smith *et al.*, "Fluorescence-guided cancer surgery—a new paradigm," *Journal of Surgical Oncology*, vol. 123, no. 8, pp. 1679–1698, 2021.

[17] M. Avanzo, L. Wei, J. Stancanello, M. Vallieres, A. Rao, O. Morin, S. A. Mattonen, and I. El Naqa, "Machine and deep learning methods for radiomics," *Medical physics*, vol. 47, no. 5, pp. e185–e202, 2020.

[18] O. Dalmaz, M. Yurt, and T. Çukur, "Resvit: residual vision transformers for multimodal medical image synthesis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022.

[19] B. Shen, Z. Zhang, X. Shi, C. Cao, Z. Zhang, Z. Hu, N. Ji, and J. Tian, "Real-time intraoperative glioma diagnosis using fluorescence imaging and deep convolutional neural networks," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 48, no. 11, pp. 3482–3492, 2021.

[20] R. A. Cahill, D. F. O'Shea, M. Khan, H. Khokhar, J. P. Epperlein, P. Mac Aonghusa, R. Nair, and S. Zhuk, "Artificial intelligence indocyanine green (icg) perfusion for colorectal cancer intra-operative tissue classification," *British Journal of Surgery*, vol. 108, no. 1, pp. 5–9, 2021.

[21] S. Zhuk, J. P. Epperlein, R. Nair, S. Tirupathi, P. Mac Aonghusa, D. F. O'Shea, and R. Cahill, "Perfusion quantification from endoscopic videos: learning to read tumor signatures," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International*

Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. Springer, 2020, pp. 711–721.

[22] S.-H. Park, H.-M. Park, K.-R. Baek, H.-M. Ahn, I. Y. Lee, and G. M. Son, "Artificial intelligence based real-time microcirculation analysis system for laparoscopic colorectal surgery," *World Journal of Gastroenterology*, vol. 26, no. 44, p. 6945, 2020.

[23] Z. Ma, F. Wang, W. Wang, Y. Zhong, and H. Dai, "Deep learning for in vivo near-infrared imaging," *Proceedings of the National Academy of Sciences*, vol. 118, no. 1, p. e2021446118, 2021.

[24] A. Xiao, B. Shen, X. Shi, Z. Zhang, Z. Zhang, J. Tian, N. Ji, and Z. Hu, "Intraoperative glioma grading using neural architecture search and multi-modal imaging," *IEEE transactions on medical imaging*, vol. 41, no. 10, pp. 2570–2581, 2022.

[25] W. Steup, Y. Moriya, and C. Van de Velde, "Patterns of lymphatic spread in rectal cancer. a topographical analysis on lymph node metastases," *European Journal of Cancer*, vol. 38, no. 7, pp. 911–918, 2002.

[26] J. Kang, Y. J. Choi, I.-k. Kim, H. S. Lee, H. Kim, S. H. Baik, N. K. Kim, and K. Y. Lee, "Lasso-based machine learning algorithm for prediction of lymph node metastasis in t1 colorectal cancer," *Cancer Research and Treatment: Official Journal of Korean Cancer Association*, vol. 53, no. 3, pp. 773–783, 2021.

[27] S.-e. Kudo, K. Ichimasa, B. Villard, Y. Mori, M. Misawa, S. Saito, K. Hotta, Y. Saito, T. Matsuda, K. Yamada *et al.*, "Artificial intelligence system to determine risk of t1 colorectal cancer metastasis to lymph node," *Gastroenterology*, vol. 160, no. 4, pp. 1075–1084, 2021.

[28] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in neural information processing systems*, vol. 32, 2019.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[32] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[33] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.

[34] S. Qian, C. Ning, and Y. Hu, "Mobilenetv3 for image classification," in *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*. IEEE, 2021, pp. 490–497.

[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.